北海道大学工学部研究報告
第108号（昭和57年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 108 (1982)

研究者向きローマ字漢字変換方式日本語処理システム

斉藤　康*　岡沢好高　栃内香次　永田邦一

（昭和56年12月26日受理）

Researcher-Oriented Japanese Word-Processing System Using Roman-Kanji Translation Method

Yasushi SAITOH, Yositaka OKAZAWA, Koji TOCHINAI
and Kuniichi NAGATA

(Received December 26, 1981)

Abstract

This paper describes a Japanese word-processing system which enables a research worker to write technical documents in his field with comparative ease.

The computer processing of Japanese sentences consisting of Kanji and Kana characters has many difficulties. Especially the input of Japanese texts has wide variety of problems, hence there exists no surpass method.

To obtain a method proper for technical writing, we investigated Kanji words and Kanji characters used in technical papers in a specific field, and developed an experimental system by means of a Romaji-Kanji translation method based on the results of the investigation. In this method, the dictionary for the translation is small and provided to each user. Moreover, its contents changes gradually and is kept at optimum for the user.

The result of several inputting experiments indicates high speed and easiness of the inputting operation, and it is concluded that the method is useful for technical writing.

1. はじめに

計算機は数値を計算する機械として誕生して，しだいにその処理対象の範囲を広げ，文字，文章を含む非数値処理にまで拡大してきた。そのひとつとして文章の入力，編集，出力を行うワードプロセッシングがあり，欧米では広く利用されている。しかし，我が国においては，日本語のもつ表記上の特徴のため文字や文章の処理は複雑になり，欧米のワードプロセッサ程度の使い勝手をもつ機器，方式では確立されていない。

日本語を計算機で取り扱う場合，特に問題となる特徴として次のようなものがある。

(1) 日本語文は漢字，ひらがな，カタカナなどを混用する漢字かな混じり文である。

(2) 漢字の種類が膨大な数である。

(3) 字体が複雑である。

電子工学科　電子機器工学講座
*現在，日本電信電話公社勤務

(4) 同音語（字）が生じやすい。

これらは日本語文の最大の特徴である漢字の存在から生じたもので，日本語の計算機処理においては漢字をどう取り扱うかが極めて重要なポイントであることがわかる。

一般に情報を計算機で処理する場合，大きく分けて「入力」，「処理」，「出力」の3つの過程に区切ることができる。漢字を含めた日本語文の処理において，「処理」の過程は主に英数字のみを扱う一般のデータ処理の延長としてほぼ問題なく扱うことができる。しかし，「入力」と「出力」に関しては上述の特徴が直接，影響してくるため日本語処理特有の手法が必要となる。このうち「出力」に関してはいわゆる漢字プリンタが出現し，当初は従来の英数字用の出力機器にくらべ，コスト，速度などの問題があったが，最近のハードウェア技術の進歩により実用上ほぼそれらの問題は解決できたといえる[1]。

一方，「入力」に関しては未解決の問題が多い。「出力」が機械から出た情報を人間が認識することであるのに対し，「入力」はその逆の過程であり，ハードウェア技術の進歩のみでの問題解決はむずかしく，人間工学的な要素も関係してくる。このため現在，多くの入力方式が提案されてはいるが，英語における英文タイプライタのような決定的なものはまだない。各方式はそれぞれの特徴によって長所がある反面，不都合な点もあり，対象とする入力作業の目的や条件に合わせて使い分けるといった現状である[2]。

かな（ローマ字）漢字変換方式はいくつかある入力方式の中でも有力な方式のひとつであり，次のような利点をもつ。

(1) 広く普及しているタイプライタ型端末を入力装置として使用できる。

(2) 特別な訓練の必要なしでも使える。

(3) 非熟練者でもある程度の入力速度が得られ，熟練すると相当高速度に達する。

一方，問題点として次のような項目があげられる。

(1) 膨大な種類の文字や漢字語を扱うため，変換の際，参照する辞書が大型のものになる。

(2) 入力をヨミによって行うため，同音語（字）が発生することを避けられない。

(3) 入力した文字列のうち，どの部分を漢字に変換するか決定し，抽出する必要がある。

これらの問題点を解決し，入力システムを具体的に構成する手法はさまざまであり，適用分野によって各々異なる。

本稿では，研究者が自分の専門分野に関する論文等を自ら作成するのに適した日本語処理システムについて報告する。このシステムは，専門分野を限定することにより，使用される語が減少し，小規模で効率的な処理が可能であることを利用するものである。以下，第2章では，ひとつの専門分野で使われる漢字語の種類，数，同音語などについての調査結果を示し，第3章では，この結果から導かれる研究者向きローマ字漢字変換方式について述べる。第4章ではそれに基づいて試作した日本語処理システムについて，また第5章ではこのシステムで採用した入力形式について述べる。

2. 特定分野における漢字語使用頻度

本方式は，特定の専門分野を限ることにより，必要な漢字語の種類が減少するという予想を基礎としている。そこでシステムを実現するのに先だち，ひとつの専門分野の論文等で文章に現われる漢字や漢字語の使われ方の傾向を知るための調査を行った。専門分野の典型として「情報処理」を選び，資料として情報処理学会誌及び同論文誌掲載の論文，解説など30編を使用した。資

料のページ数，漢字語語種などの一覧を表１に示す。

表１　資　料　一　覧

全体	資　料　数	30編
	総ページ数	189ページ
	漢字語語種	3,671語
	漢字語のべ語数	34,906語
頻度2以上	漢字語語種	2,293語
	漢字語のべ語数	33,528語
	全体に対する割合	96.1%
	漢字の種類	786字

図１は漢字語ののべ語数を累積していったときの漢字語語種の増加の様子を示している。語種の増加率はのべ語数の累積につれて減少し，のべ語数25,000以上ではほぼ一定勾配の直線となっている。この結果から，のべ語数が25,000語ぐらいまでで各資料で共通的に使われる漢字語はほぼ出つくし，それ以降，増加する漢字語は資料ごとに散発的に使われる語と考えられる。したがって，語種が極めて多くなるまでこの一定増加傾向が続くと予想される。

図２は各資料において，その資料で初めて出現した漢字語（新出語）のその資料中ののべ語数に対する割合が資料の累積につれてどのように変化するかを示したものである。これから新出語の比率はのべ語数が25,000語以上ではほぼ一定の範囲内で変動していることがわかる。図１の場合と同様，この傾向は資料数を増加しても続くと考えられる。

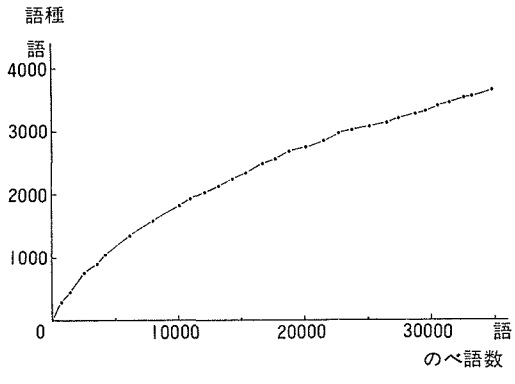

図１　漢字語語種の累積変化

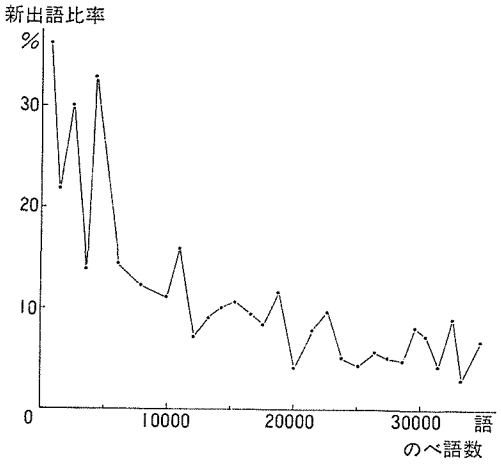

図２　新出語比率の推移

図３は使われた漢字語を使用頻度別に区分し，その分布を全語種に対する比率で示したものである。漢字語の語種数でみると頻度の少ないものほど割合が大きくなっており，頻度１の語で４割近くを占める。逆にいうと，出現した漢字語のうち４割は調査した30資料中で１度しか使われ

ていないことになる。このように頻度の少ない部分の語が前述した各資料で散発的に使われる漢字語であると考えられる。しかし，のべ語数でみると頻度の少ないものの割合はずっと少なくなり，ピークは頻度33～64の位置にくる。このピークの位置は調査資料数を増加してゆけば，しだいに頻度の大きい方向にずれてゆくと考えられる。この結果から頻度1の語ののべ語数に対する割合はわずか4％程度となっており，頻度2以上の語でのべ語数の96％を占めていることになる。このことから頻度の少ない語を除くことにより，語種数をかなり減少させることができ，しかもカバー率の減少をわずかに抑えることができて，変換用辞書の小型化が可能であることがわかる。

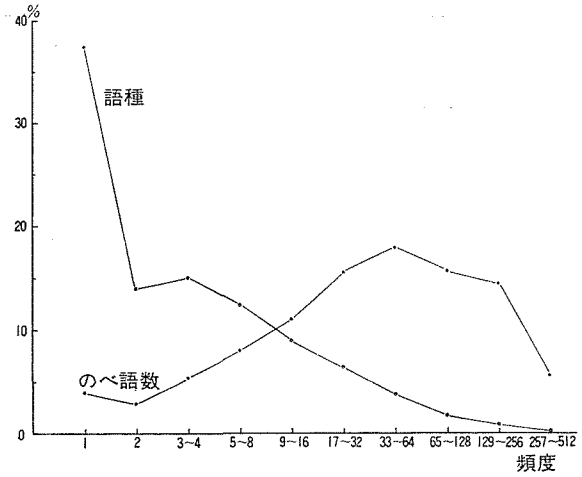

図3 語種，のべ語数の頻度別比率

図4は資料を累積して漢字語語種が増加していく過程で，全体ののべ語数のうち同音語をもつ語の占める割合の変化を示したものである。これは，変換用辞書に収容する漢字語の語数によって，変換の際，どのくらいの比率で同音語が発生するかを示している。語種の増加につれて同音語の発生率はしだいに増加し，語種数が約3,700語のところで20％となっている。この比率は語種数が増加すればますます増加すると予想され，辞書に収容する語数をあまり多くすることは，同音語発生の面から適当でないことになる。

以上の結果から，散発的に出現する新出語を処理する適当な方式を組込むことにより，少語数の辞書で十分有効な変換が可能であると結論される[3]。

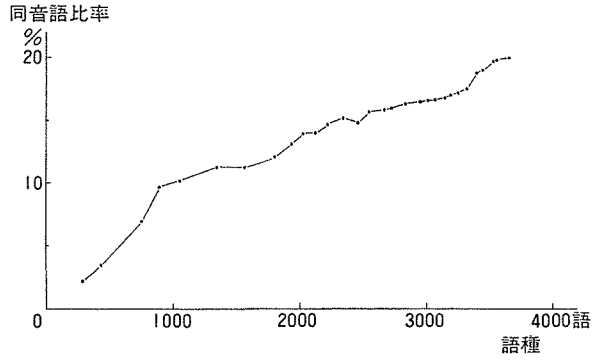

図4 同音語の占める比率

3. 研究者向き日本語処理システム

3.1 入力方式

研究者が論文など専門分野に関する文書を自分で容易に作成するという目的に適した入力方式を決定するにあたって考慮すべき条件として，次の諸点があげられる。

(1) 入力作業については素人である研究者が入力する。

(2) 一般に研究者は TSS 端末など欧文タイプライタ型端末の操作に慣れており，また，この種の端末が研究室には広く普及している。

(3) 作成される文書の範囲はある専門分野に限られる。

このような条件から最適な方式としてローマ字漢字変換方式を採用した。この方式は語のヨミのとおり入力するため素人でも比較的高速な入力速度が得られ，また，研究者にはローマ字入力の方がカナ入力より容易に入力できる。変換を漢字1文字ごとに行うか，複数の文字からなる漢字語単位で行うかの2種類の方法が考えられるが，前者では変換用の辞書が小さくてすむが，同音字の発生が多くなり，逆に後者では，変換用辞書は大きくなるが同音語の発生を少なく抑えることができる。本システムでは上記(3)の条件から辞書の語数を比較的少数にすることが可能であるため，後者の語単位ローマ字漢字変換方式を採用した。

また，漢字に変換する部分の抽出については自動的に行う方式と入力者が陽に指定する方式が考えられるが，本来ある文字を漢字，かなのどちらで表わすかは文章の作成者の意志によるものであり，特に本システムの場合，入力者が文章の作成者自身であって打鍵時にどの部分を漢字で表わすかを知っていることから，入力時に漢字変換部分を陽に指定する方式とした[4]。

3.2 変換用辞書

ローマ字漢字変換方式では，入力されたローマ字列を漢字かな混じり文に変換する際に参照する変換用辞書が重要な役割を果たす。この辞書の内容により変換率，同音語発生率などが決定づけられる。また，本システムで採用した語単位の変換では文字単位の変換にくらべて辞書容量が大きくなる欠点を有し，したがって，いかにして辞書を小型化し適当な容量に抑えるかが重要になる。

本システムでは研究者が自分の専門分野に関する文書を作成することを目的としており，広範囲の文章を扱う場合にくらべて使われる漢字語の種類は少なくてすむ。すなわち，辞書に登録すべき語数が少なくてすむ。このことは第2章の結果からも明らかであり，資料全体で現われた漢字語3,671語のうち頻度2以上の語2,293語でのべ語数の96.1%を占めており，広い範囲の文書を対象とした市販のワードプロセッサが数万語の辞書を備えているのと比較して大きな差がある[5]。また，この傾向は対象を1人の著者の文章にまでしぼることにより顕著になる[6]。そこで本システムでは変換用辞書を小容量にして使用者個人ごとに保有することとした。辞書の容量は前記，第2章の結果から2,500語とした。

変換用辞書の内容は使用者各個人に適したものでなくてはよい変換結果を得ることはできない。また，容量を小型に抑えるため，なおさら収容する漢字語は厳選し，辞書を最適なものにする必要がある。しかし，最初から各個人にとって最適な内容の辞書を得ることは困難であり，最適な内容も時間とともに変化してゆくと考えられる。そこで本システムでは辞書を，内容が使用につれて変化し，最適な状態を保つ動的な構造とした。すなわち，辞書に登録されていない漢字語を変換しようとする場合はその語を辞書に新たに登録することになるが，その際，登録してあ

る漢字語の中から，使用頻度が少なく，長い期間使われていない語を選び，その語を削除して領域をあけ，そのあとに新規の登録を行う。この操作の繰り返しにより辞書の内容は使用者にとって最適な状態を保つことができる[7]。

4. システム構成

4.1 全体の構成

本システムは北大大型計算機センターの HITAC システム上で実現している。その概略を図5に示す。

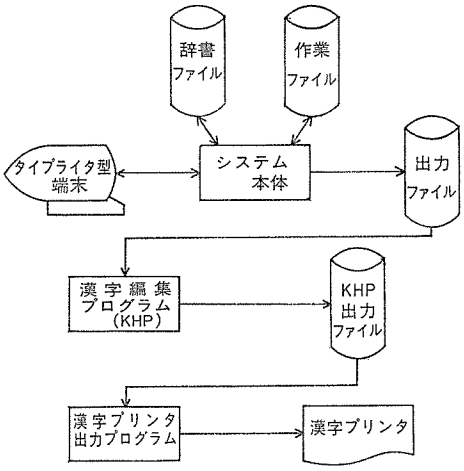

図5 システム構成

このシステムでは入力は欧文タイプライタ型端末，すなわち，漢字表示のできない通常の TSS 端末から行う。入力されたローマ字列を漢字符号に変換した結果は，出力ファイルに書き込まれる。そのファイルは HITAC システムに既設の漢字編集プログラム (KHP) の入力ファイルとなり，行単位，ページ単位の編集が行われ，最終的に漢字プリンタ出力プログラムを経て，漢字プリンタから漢字かな混じり文として出力される。

なお，処理の前半は会話型処理で行うが，後半 (KHP 以降) は KHP の仕様によりバッチ処理で行う。

また，所望の文書が一度の入力作業で完成することはまれであり，何度か入力ローマ字列を修正してから再入力するのが普通である。そこで本システムでは端末から入力されたローマ字列を作業ファイルに記録しておき，必要があればこのファイル上で変更個所を修正し，再入力できるようにした。

4.2 変換用辞書の構成

変換用辞書の内容の例を図6に示す。検索時間を短縮するため辞書内で各漢字語は使用頻度の多い順に並べられている。したがって，漢字語が使用され頻度が増加すれば語の配列順序は変わることになる。

収録されている項目は，語番号，漢字語のヨミのローマ字つづり，頻度などのほかに，その語が最後に使われてからどのくらいの期間が経たかを示す履歴カウンタが設けられている。このカ

ウンタの値はシステム使用に際し，セッションが開かれる度に1ずつ値が増え，その漢字語が使用されると値が0になる。したがって，この値が大きいものほど長い期間使われていないことになり，3.2で述べたように新しい語を登録するために辞書の領域を1語分あける必要が生じたとき，頻度とこのカウンタの値を参照して，今後最も使われそうにない語，すなわち，頻度が少なく履歴カウンタの値が大きい語を選び，削除する[8]。

また，本システムでは漢字表示のできない端末から入力することを考えて，同音語が発生した場合などに各漢字語の特定ができるような補助情報を辞書に記録している。補助情報としてはその漢字語に対応する英単語や送りがなを含めたローマ字つづりなどを使っている[7]。

語番号	版番号	履歴カウンタ	ローマ字つづり漢字語	漢字符号	頻度	同音語数	補助情報
24	1	8	Zikan	BBFEB4D6	165	1	TIME
1,745	1	14	Naga	C4B9	18	2	NAGA-I

図6 変換用辞書の内容

4.3 処理の流れ

端末から入力されたローマ字列は文字種ごとに分離される。そのうち漢字部分は漢字語単位に変換用辞書内の語と比較され，唯一つの語と一致したときはその漢字符号を出力する。また，同音語が発生した場合，すなわち辞書中の複数の語とヨミが一致した場合は端末にそれらの補助情報をすべて表示し，入力者に選択させ，選ばれた語の漢字符号を出力する。また，入力された漢字語が辞書に登録されていなかった場合は同様にその旨表示し，その語に関する情報の登録を要求する。

漢字部分以外のひらがな，カタカナ，数字，記号などの部分は入力されたローマ字列と漢字符号が各々1対1に対応するので，その符号を出力する。

以上のようにして漢字符号が書き込まれた出力ファイルが作成され，編集処理を行った後，一括して漢字プリンタから漢字かな混じり文として出力される。

5. 入力形式

5.1 ローマ字表記法

システムへの入力は欧文タイプライタ型端末からローマ字つづりで行うが，一般に使用されているローマ字表記法では機能的に不足または不都合な点がある。そこで本システムではこれと矛盾しないようないくつかの規則を付け加えた独自の表記法を用いることとした。その主な特徴を以下に示す。

(1) 外来語なども表現できるように下つき小字ァ，ィ，ゥ，ェ，ォ，ャ，ュ，ョ，ッを含めたかな文字表記すべてについて対応するローマ字表現を定めた。

(2) かなは小文字で表わす。

例 hiragana （ひらがな）

(3) 促音「っ」は「q」で表わす。

例 kiqto （きっと）

(4) 漢字語はその先頭文字を大文字で表わす。なお，漢字語の後にかなが続くときは区切り

として間に空白をひとつ置く。

例 NihongoSyori （日本語処理）

Kanzi␣tokana （漢字とかな）

(5) カタカナの部分は大文字の X で前後をかこんで表わす。

例 SyoriXsisutemuX （処理システム）

(6) ローマ字のまま出力するときは「&␣」と「&」でかこむ。

例 &␣TSS&Tanmatu （TSS 端末）

(7) 数字やキーボード上にある記号はそのまま使う。

例 Dai1Syou （第1章）

(8) 漢字プリンタにあってキーボード上にない文字，記号については各々の文字，記号に対して特定の略記号名を定め，それを「&」でかこんで表わす。

例 6&dv&3=2 （6÷3=2）

(9) 改行や文字の大きさなどを指示する制御記号も各々について略記号名を定め，「&.」と「&」でかこむ。

例 Kaigyou␣site&.nl1&Insatu␣suru. （改行して印刷する。）

以上によりローマ字によって日本語文を表現することができる。入力文とその変換結果の漢字かな混じり文の例を図7に示す。

入 力 **XsisutemuXhenoNyuuryoku haTuuzyou no& TSS&Tanmatu yori0kona u.**

出 力 システムへの入力は通常のＴＳＳ端末より行う

図7 入 出 力 例

5.2 同音語及び新規登録処理

システムへの入力中に同音語や辞書中に登録されていない漢字語が現われると，同音語処理や新規登録処理を行う必要がある。

同音語が発生するとシステムは辞書中に存在する同音語すべての補助情報を端末に表示して入力者に選択を促す。そこで入力者は所望の漢字語を番号で選ぶ。なお，ひとつの文書内では同音語中の特定の語のみを多く使うことがよくあるため，一時的に同音語をその語に固定しておき，以後の同音語選択の手間を省くことができる。また，表示された同音語中に所望の語がない場合はその語を新規登録処理で辞書に登録することになる。同音語選択処理の例を図8に示す。

```
:
ZyouhouSyori

The word 'Zyouhou' has following 'Douongo'. Select a number!
 1   INFORMATION
 2   KAKEZAN
If another,input '0'!
:
1
```

図8 同音語選択処理（下線は入力部）

辞書にない未登録語が入力されるとシステムはその旨を表示し，入力者に新規登録処理を促

```
:
Keisanki woTuka i,

Not found in the dictionary : Keisanki
Select a number!
 (1) Write into the dictionary
 (2) Reenter
 (3) Print 'Geta'
:
1

Input for the new word:Keisanki
 Kanzi codes:B7D7BBBBB5A1
 Information:COMPUTER
```

図9 新規登録処理（下線は入力部）

す。入力者はコードブックから調べた漢字符号とその漢字語を識別できるような補助情報を入力する。新規登録処理の例を図9に示す。

6. おわりに

本報告では研究者向き日本語処理システムについて述べた。その特徴を要約すると以下のようになる。

(1) 入力方式は入力作業には素人である研究者でも容易に扱え，広く普及している欧文タイプライタ型端末が使えるローマ字漢字変換方式を採用した。

(2) 同音語の発生を抑えるため，変換の単位は漢字語とした。

(3) 変換用辞書の容量を小型化し，同音語の発生を抑えるため，辞書は使用者個人ごとに保有することとし，収録する語は特定専門分野に限定した。

(4) 辞書の内容は使用されるにつれて変化し，使用者にとって最適な状態を保つよう動的な構造とした。

(5) 漢字とかなの区別は入力者の指定によって決定する方式とした。

(6) 各漢字語に対する補助情報の付加により，漢字表示のできない端末から同音語選択を可能にした。

このシステムの実現により，従来のタブレット形の漢字入力装置にくらべ2～3倍の入力速度が得られ，入力者の負担も大幅に軽減できた。辞書の最適化が進むにつれ，より速い入力速度が期待できる。本論文の第1章の一部をこのシステムで作成した結果を付録に示す。

しかし，同音語選択処理や未登録語の新規登録処理がスムーズな入力作業を妨げることは避けられず，入力速度に大きな影響を及ぼす。同音語選択処理に関しては表示された中から数字を選ぶ簡単な作業であり，また，一時固定化によってもある程度，簡単化できる。しかし，新規登録処理はコードブックから漢字符号を検索するという作業を伴い，これは端末操作から一旦，離れることになって入力作業の流れを大きく乱すことになる。そこでコードブックによらず，端末操作のみで新しい漢字語を登録できることが望ましい。このためにはコードブックに対応するものをシステム内に保持することが必要であるが，表1にあるように頻度2以上の漢字語2,293語で使用された漢字は約800文字にすぎず，このことから比較的少数の漢字を記録していれば必要な漢字語のうち，かなりの割合の語を構成できると考えられる。そこで，1,000文字程度の漢字とその漢字符号を収録した文字単位の辞書を設け，これを用いて必要な漢字語を合成し登録する機能を現在開発中である。この機能の組込みと，それによる入力処理効率の向上については今後あらためて報告する予定である。

謝辞 本研究を行うにあたり，漢字システムの操作法など御指導いただいた北大大型計算機センター関係各位ならびに本研究室に在籍され，卒業研究としてファイル構造の調査等に貢献された吉田裕司君に深謝致します。

参 考 文 献

1) 草野玄三：電子通信学会誌，63, No. 7, p. 712 (1980)
2) 渡辺定久：同，p. 707
3) 栃内香次，斉藤康：情報処理学会22回大会講演論文集，3 I-1 (昭56)
4) 栃内香次，高平敏男，斉藤康：情報処理学会21回大会講演論文集，7 I-9 (昭55)
5) 大倉信治：bit, 13, No. 4, p. 329 (1981)
6) 斉藤康：北大工学部卒業論文 (昭55)
7) 斉藤康，岡沢好高，栃内香次，永田邦一：電子通信学会情報・システム部門全国大会講演論文集，658 (昭56)
8) Tsichritzis, Bernstein：オペレーティングシステムの基礎，p. 135，日本コンピュータ協会 (1967)

付 録

本システムによる文書作成の例を以下に示す。

入 力

```
&.cs10,cf2,lp18,tab218,sk15&1 hazimeni&.nl2&
 Keisanki haSuuti woKeisan suruKikai tositeTanzyou
site,sidainisonoSyoriTaisyou noHan'i woHiro ge,Mozi,
Bunsyou woHuku muHisuutiSyori nimadeKakudai sitekita.
sonohitotutositeBunsyou noNyuuryoku,Hensyuu,Syuturyoku
woOkona uXwa-dopuroseqsinguXgaari,Oubei dehaHiro ku
Riyou sareteiru.sikasi,WagaKuni nioiteha,Nihongo nomo
tuHyoukiZyou noTokutyou notameMozi yaBunsyou noSyori
haHukuzatu ninari,Oubei noXwa-dopuroseqsaXTeido noTuka
iKaqte womotuKiki,Housiki haKakuritu sareteinai.&.nl1&
 Nihongo woKeisanki deToriAtuka uBaai,Toku niMondai to
naruTokutyou tositeTugi noyounamonogaaru.&.nl1&
  (1) Nihongobun haKanzi,hiragana,XkatakanaXnadowo
Kon'you suruKanzi&.nl1&
     kanaMazi riBun dearu.&.nl1&
  (2) Kanzi noSyurui gaBoudai naKazu dearu.&.nl1&
```

出 力

1 はじめに

計算機は数値を計算する機械として誕生して，しだいにその処理対象の範囲を広げ，文字，文章を含む非数値処理にまで拡大してきた．そのひとつとして文章の入力，編集，出力を行うワードプロセッシングがあり，欧米では広く利用されている．しかし，我が国においては，日本語のもつ表記上の特徴のため文字や文章の処理は複雑になり，欧米のワードプロセッサ程度の使い勝手をもつ機器，方式は確立されていない．

日本語を計算機で取り扱う場合，特に問題となる特徴として次のようなものがある．

（１） 日本語文は漢字，ひらがな，カタカナなどを混用する漢字かな混り文である．

（２） 漢字の種類が膨大な数である．